Introduction to SNPs and STRs

Single Nucleotide Polymorphisms and Short Tandem Repeats

UZH-BIO392
Feifei Xia
feifei.xia@uzh.ch

Introduction to STRs

Short Tandem Repeats

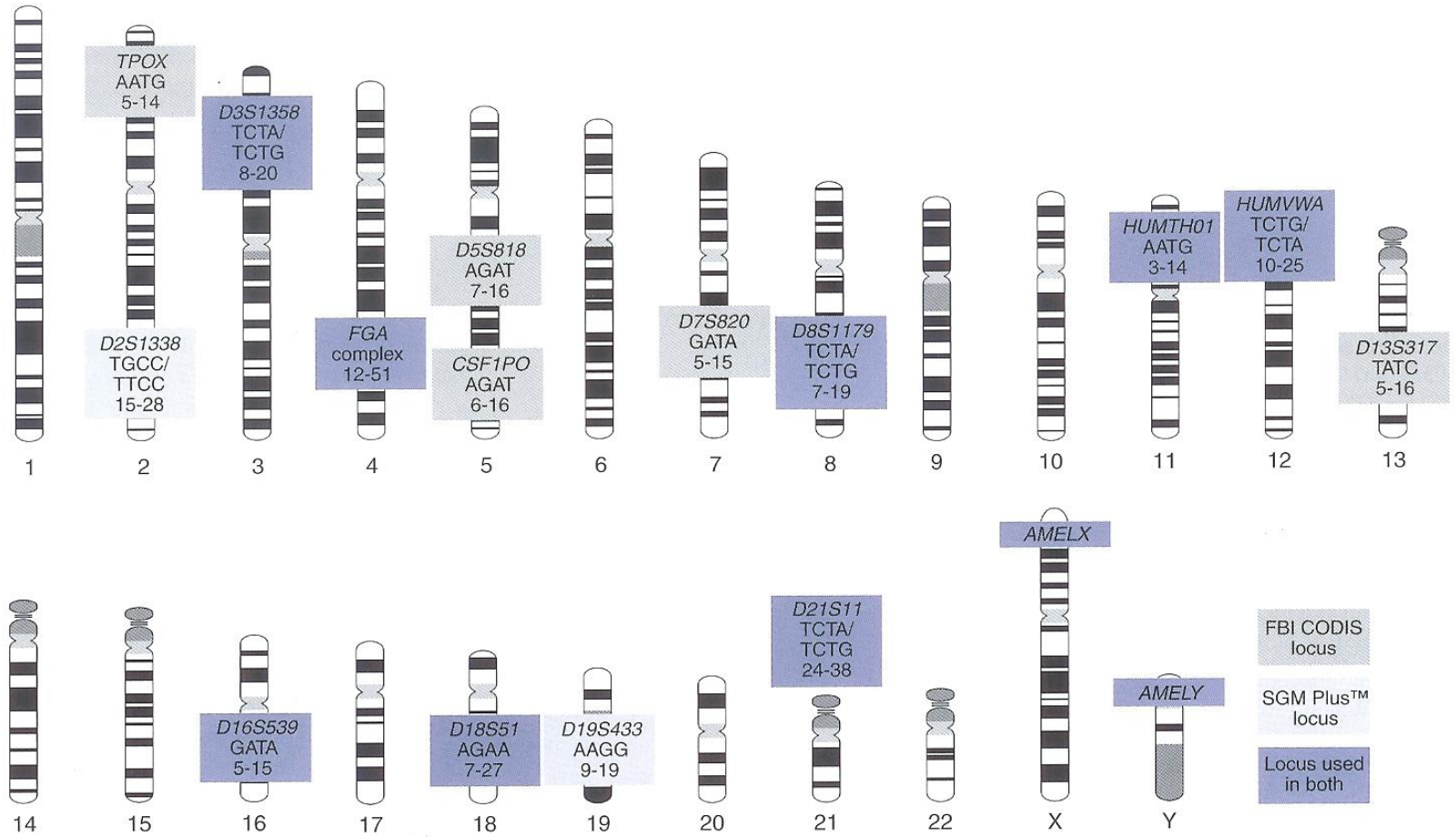
Types of Human Genetic Variants

Differences in DNA among individuals drive many types of phenotypic differences.
There are many types of genomic differences between individuals

- Structural variations (SVs)
- Copy number variations (CNVs)
- **Short tandem repeats (STRs)**
- **Single nucleotide polymorphisms (SNPs)**

STRs in forensics analysis

CODIS (USA) uses 20 core STR loci standardized by the FBI



https://www.mun.ca/biology/scarr/CODIS_&_SGM.html

What are STRs?

Repetitions of 1-6 bp DNA units

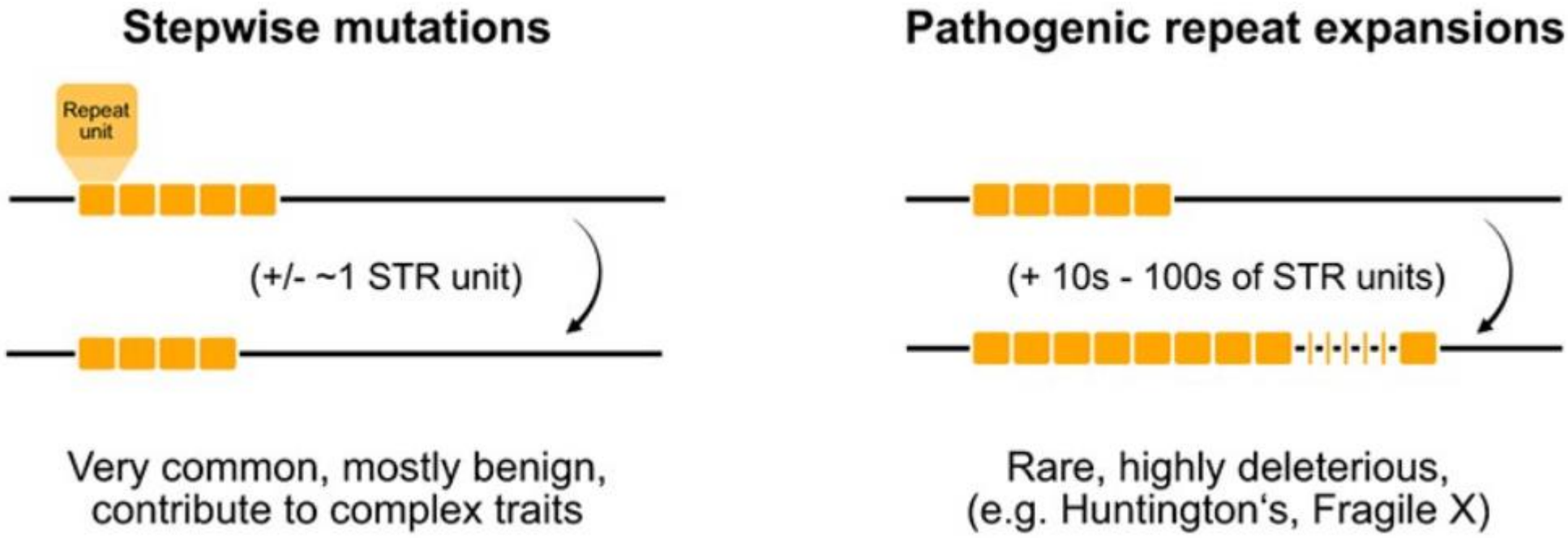
- Abundunt and polymorphic
- High variability in repeat number between individuals

Repeat unit : AT
Repeat length : 9

STRs vs SNPs – A Comparison

Feature	STRs (Short Tandem Repeats)	SNPs (Single Nucleotide Polymorphisms)
Definition	Repeating sequences of 1–6 base pairs	A single base change in the DNA sequence
Mutation Type	Insertion/deletion of repeat units	Substitution of one base for another
Mutation Rate	High (~10^{-3} per generation)	Low (~10^{-8} per generation)
Alleles per Locus	Multi-allelic (many allele lengths)	Typically biallelic (two possible alleles)
Size Range	Varies from 5 to 100+ base pairs	Just 1 base

STR length variations

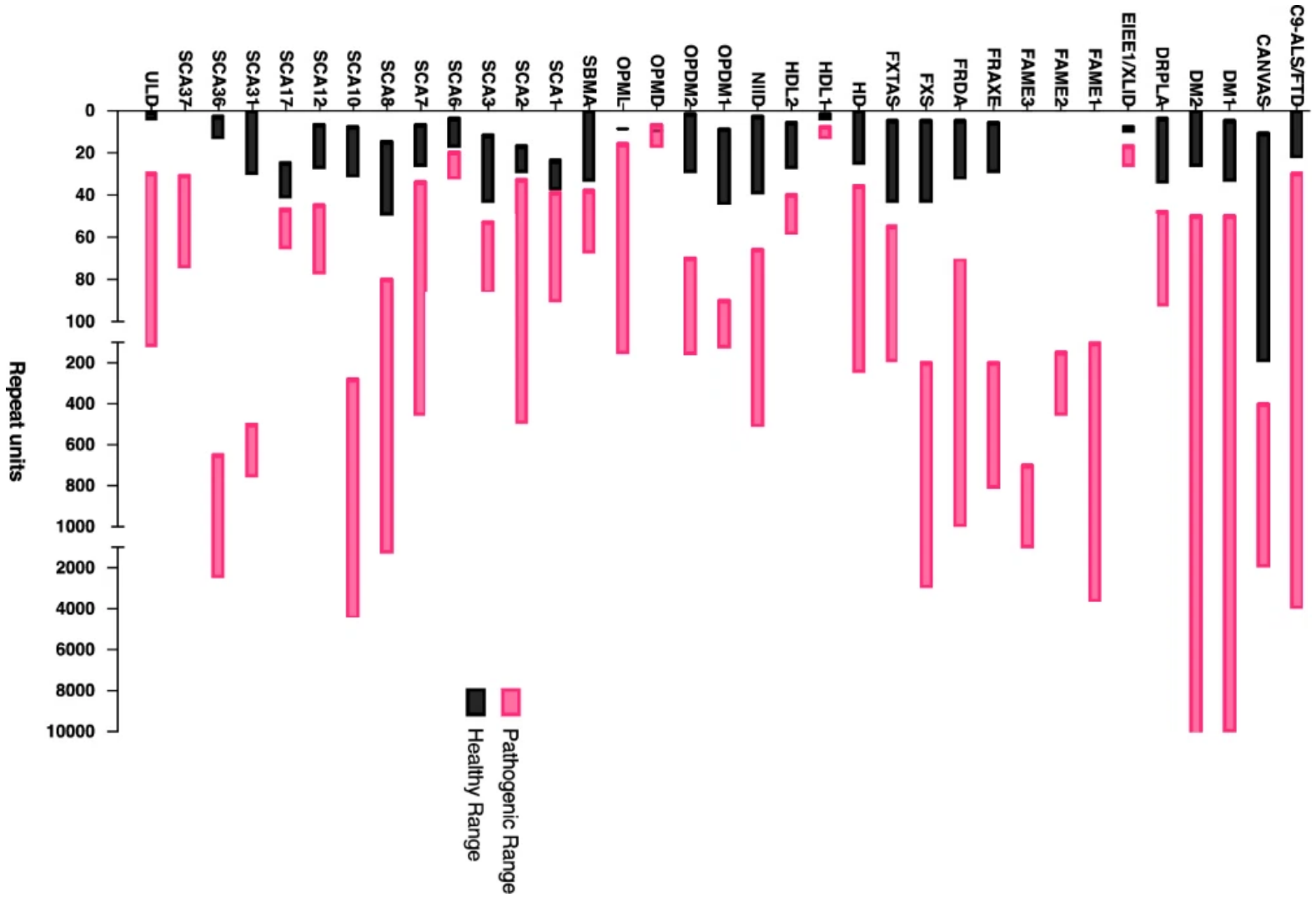


STRs in cancer and disease

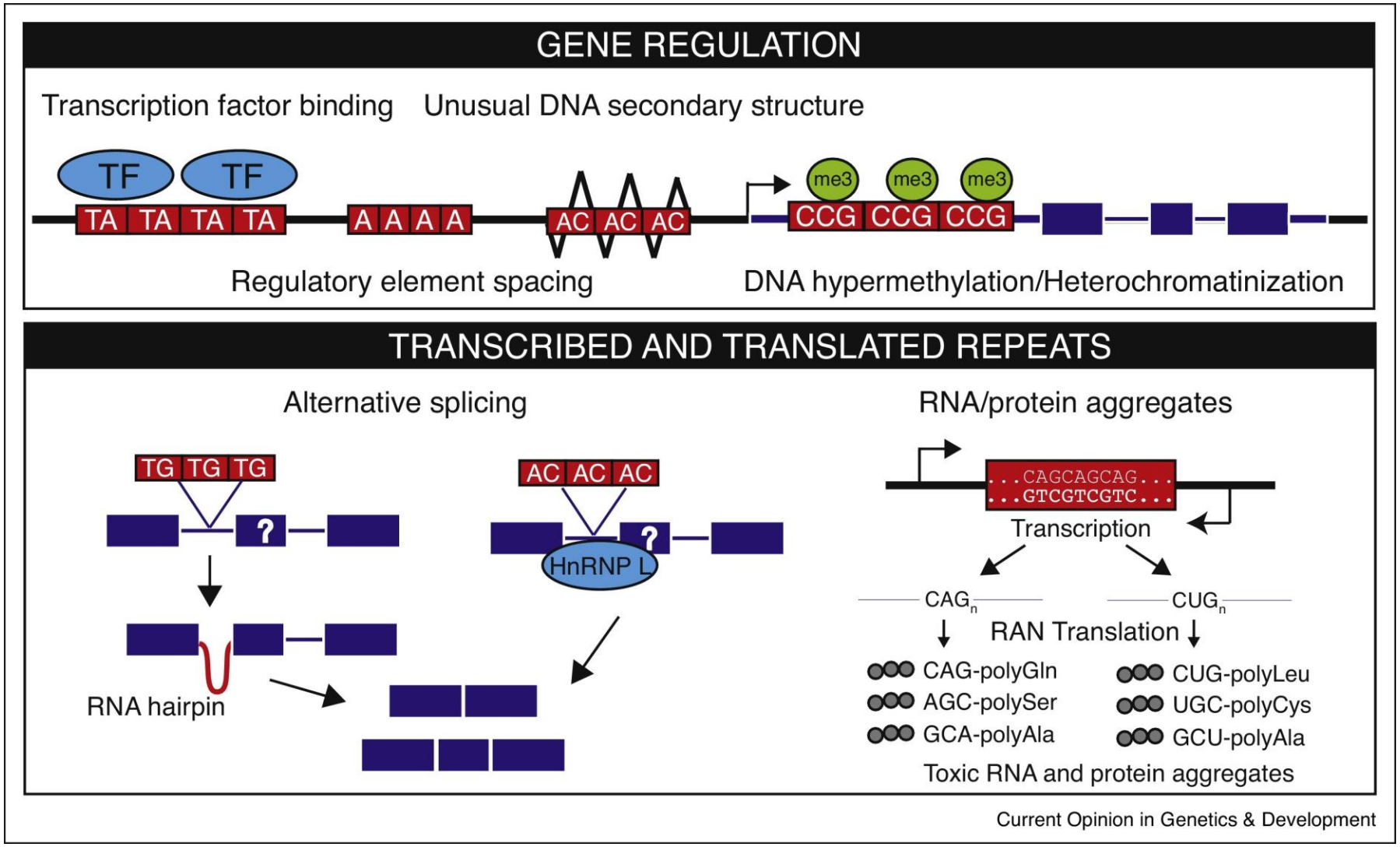
- Microsatellite instability (MSI) is the condition of having abnormally high rates of mutation in microsatellites (STRs) within the genome, caused by a defect in the DNA mismatch repair (MMR) system. MSI tumors respond better to immunotherapy.

- Pathogenic STR expansions: Huntington’s, Fragile X

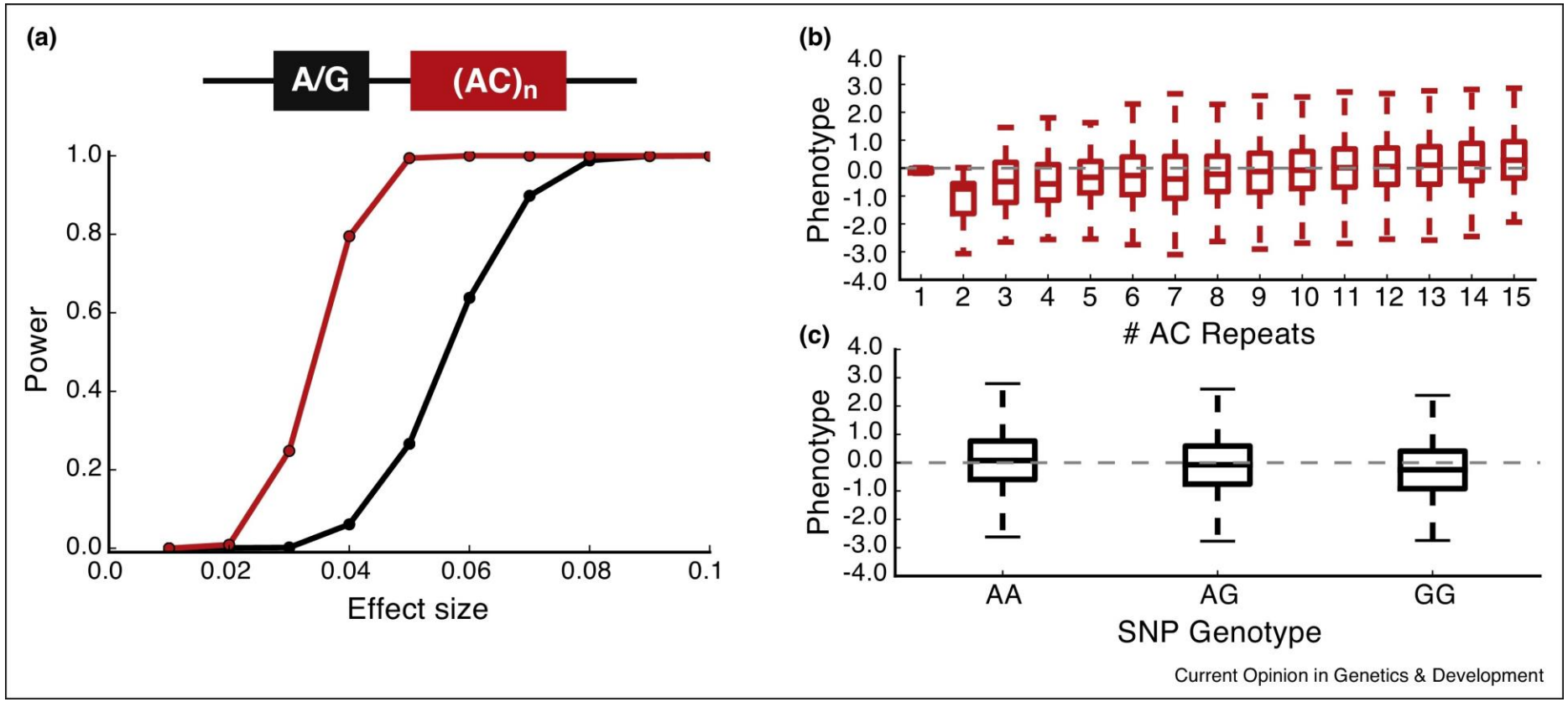
Short Tandem Repeat Expansion Disorders



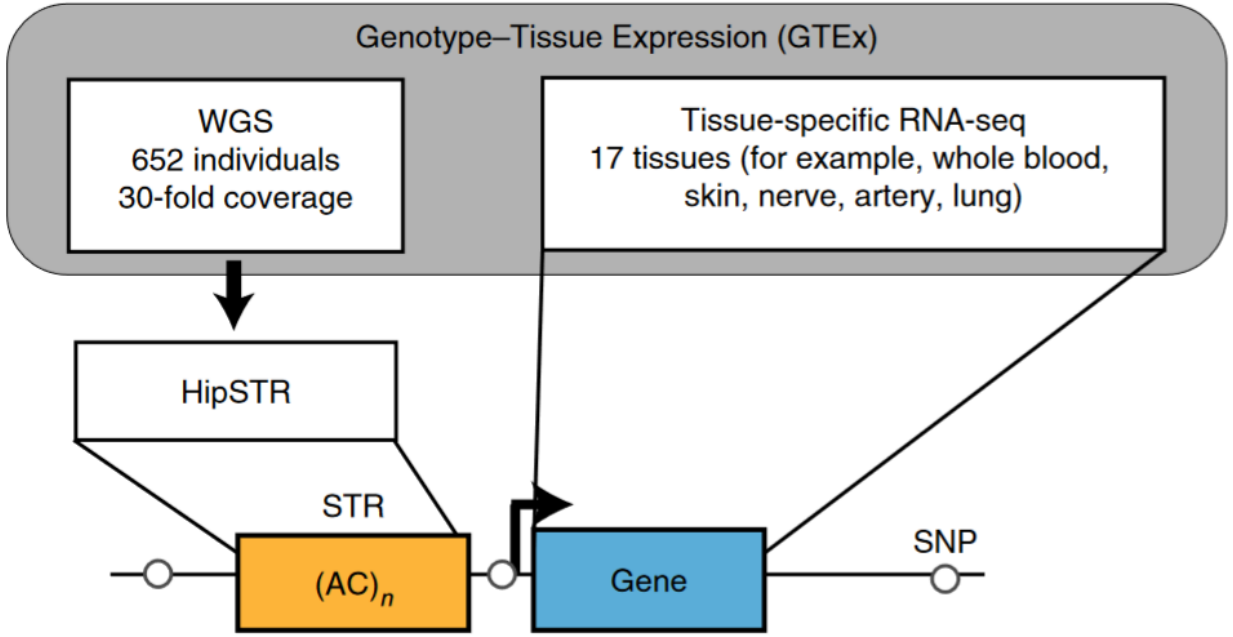
How do STRs affect gene function?



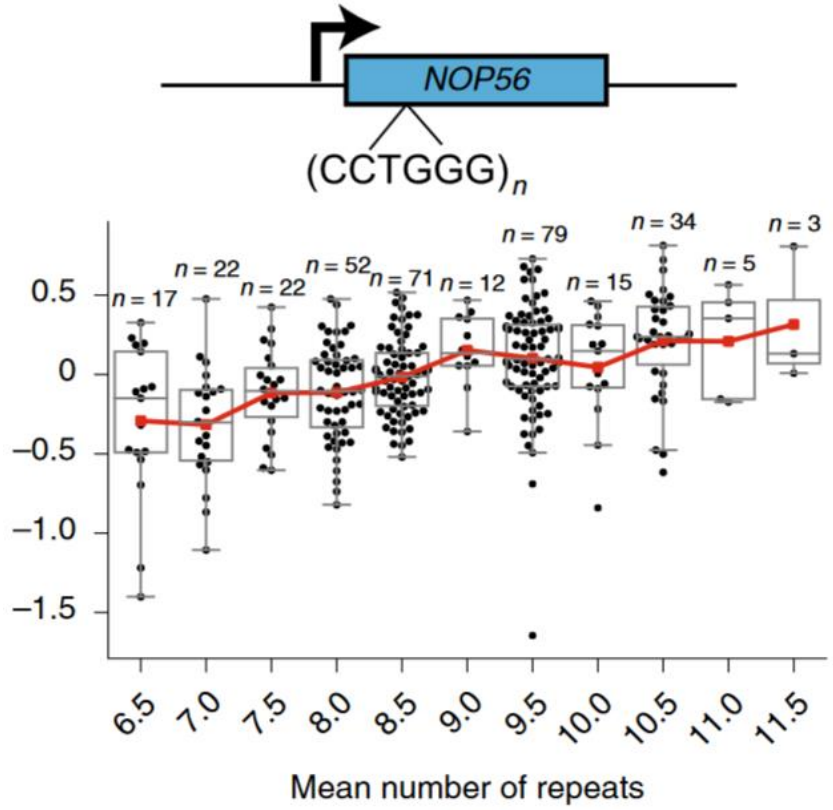
Short Tandem Repeats offer much more genotypic variation than SNPs



How do STRs affect gene expression



eSTR identification



WebSTR

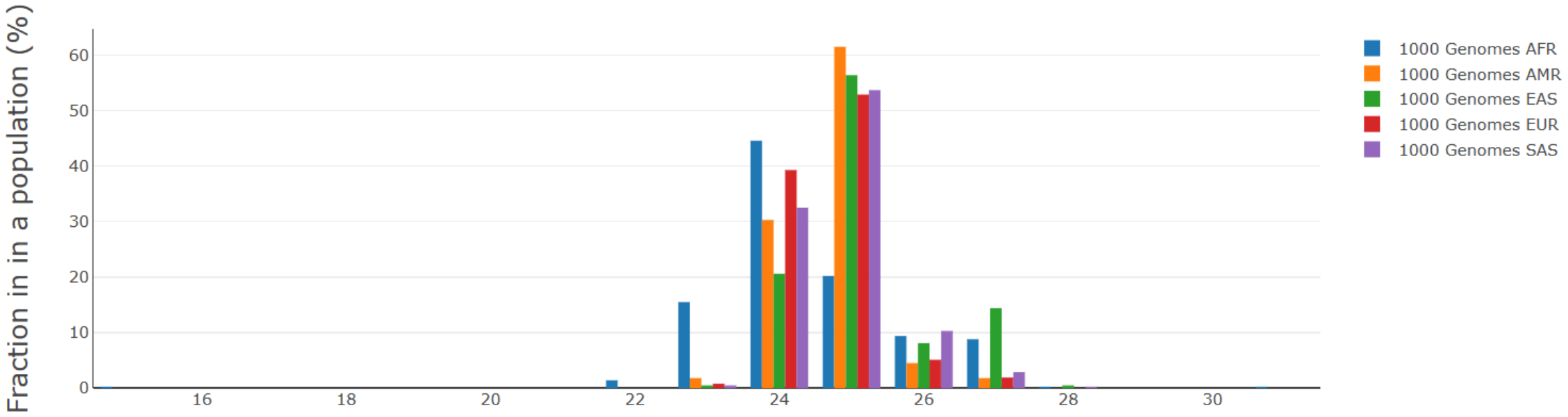
a population-wide database of short tandem repeat variation in humans

Available datasets

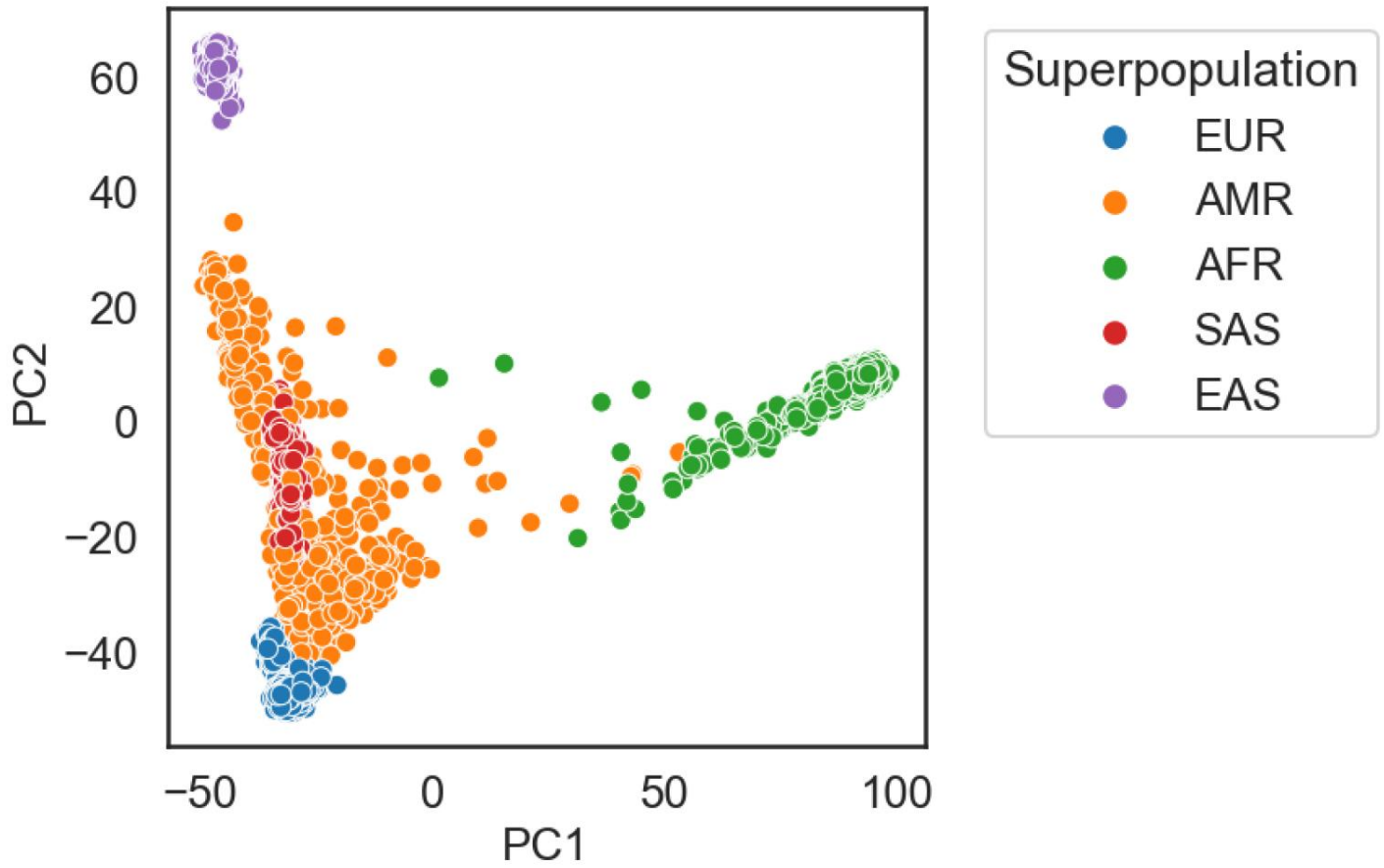
Panel alias	Human Genome Assembly Version	Annotation Method	Genotyping Methods	STRs	Cohorts	Samples	Available Data
ensembletr_hg38	GRCh38.p2	TRF	EnsembleTR ExpansionHunter GangSTR HipSTR	1,710,833 total reference panel with allele frequency data available 35,998 1,133,225 1,331,280	1000Genomes/H3Africa	3,550	Allele frequencies
gangstr_crc_hg38	GRCh38.p2	TRAL	GangSTR	1,548,993 Reference STRs 219,394 were found to have at least 1 variation call in the dataset 81,262 STRs have reliable variation data	Sinergia-CRC	412	Average variation of the number of repeat units in the dataset
hipstr_hg19	GRCh37	TRF	HipSTR	~1.6 million	GTEx SSC 1000Genomes SGDP	652 1,916 150 300	Allele frequencies,trait associations, imputation metrics, mutation parameters

WebSTR

population-wide database of short tandem repeat variation in humans



Population Structure from STRs



STR Genotyping Methods

STR genotyping refers to identifying the number of repeat units (alleles) present at specific STR loci in a DNA sample.

PCR

- The gold standard in forensics and paternity testing
- Microsatellite instability (MSI) testing in certain cancers

Genotyping from NGS data

- HipSTR: Handles PCR stutter and alignment errors
- ExpansionHunter: Designed for pathogenic repeat expansions (e.g., Huntington's disease)
- STRetch: Detects repeat expansions from WGS
- GangSTR: Efficient genotyping from large WGS datasets

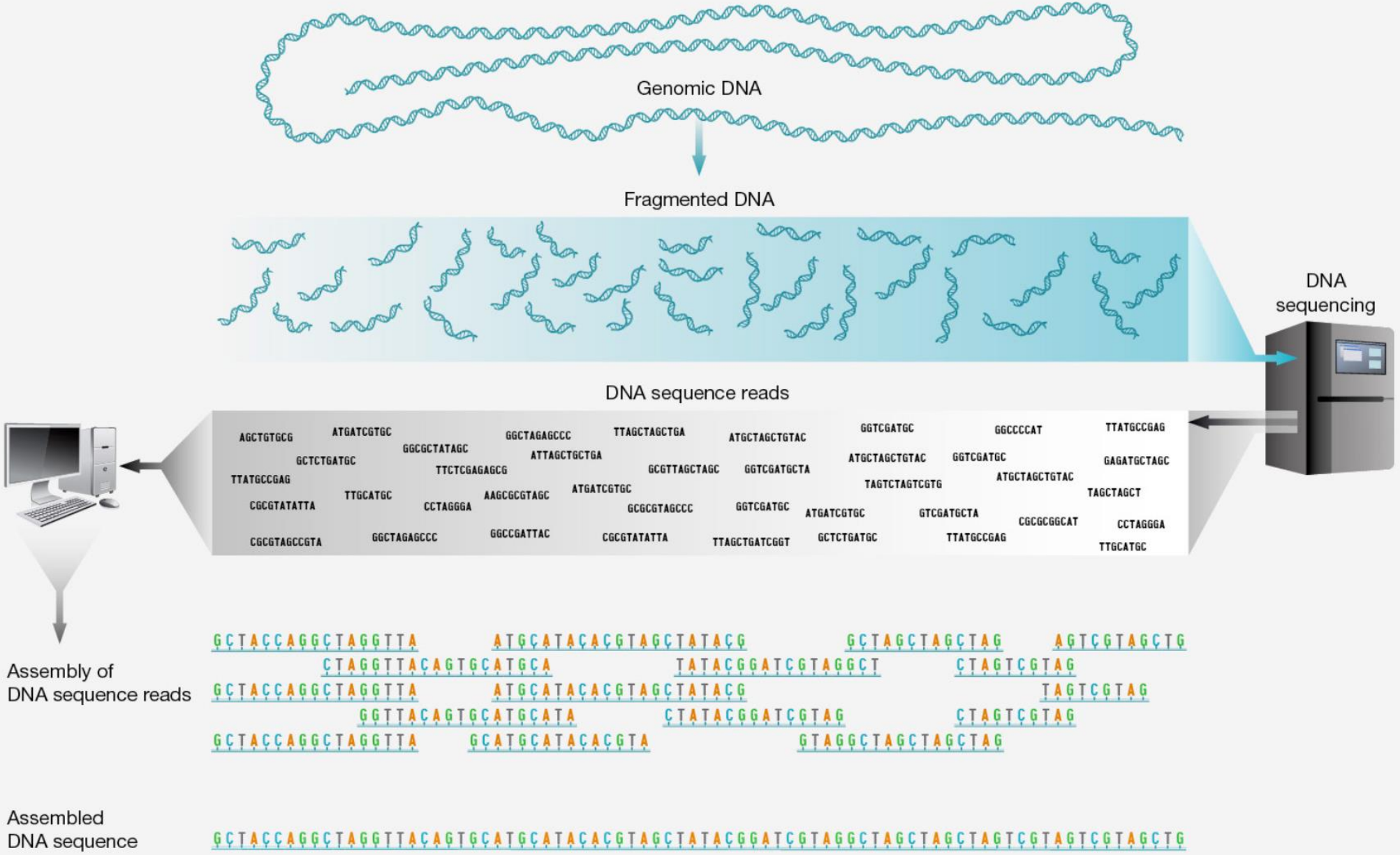
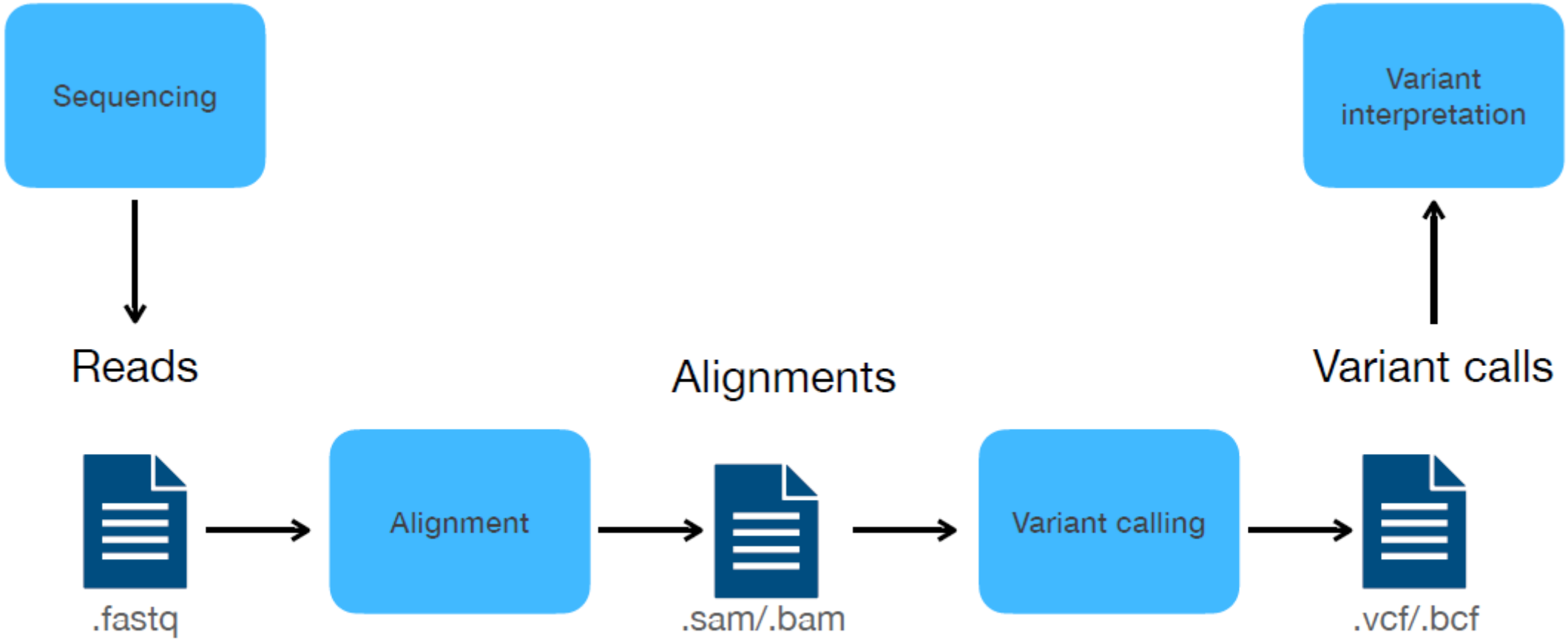


Image courtesy of National Human Genome Research Institute (https://www.genome.gov/genetics-glossary/Shotgun-Sequencing)

Bioinformatics Workflow

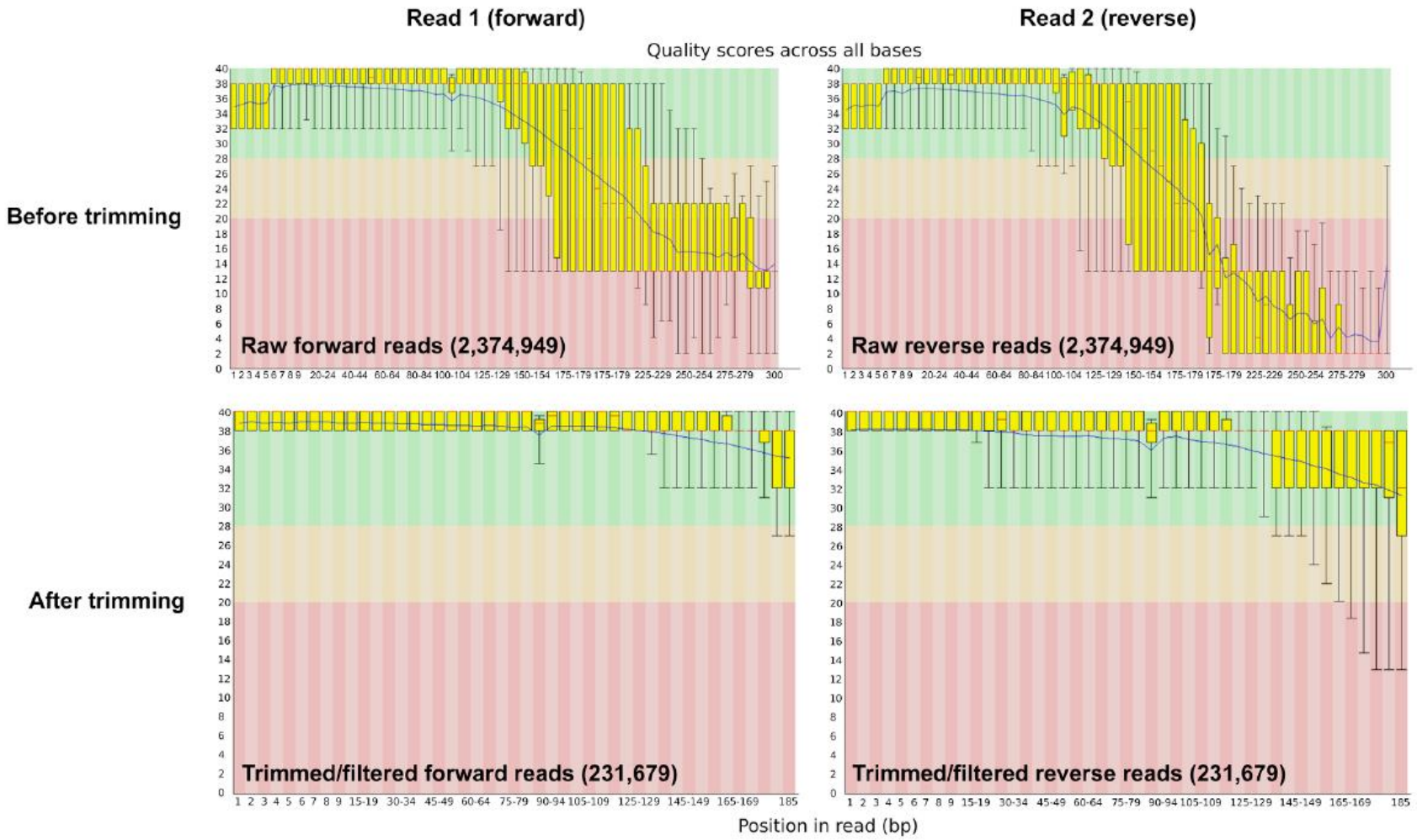


Reads are processed before analysis

Typical read processing steps:

- Adapter trimming (remove leftovers from sequencing process)
- Quality trimming (remove low Phred-score bases)
- Length filter (remove too-short reads)

Reads are processed before and after



https://www.jove.com/t/54936/metagenomic-analysis-of-silage

Short reads

.....
GGTCTGGATGC
CGGTCTGGATGC
GCGGTCTGGATG
GCGGTCTGGAT
GGCGGTCTGGAT
GGCGGTCTGGA
TCTATGCGGGCCCCT
TCTATGCGGGCCCC
ATCTATGCGGGCC
TATCTATGCGGGC
TTATCTATGCGGG
CTTATCTATGCGGG

Alignment of reads to the reference genome and SNP calling

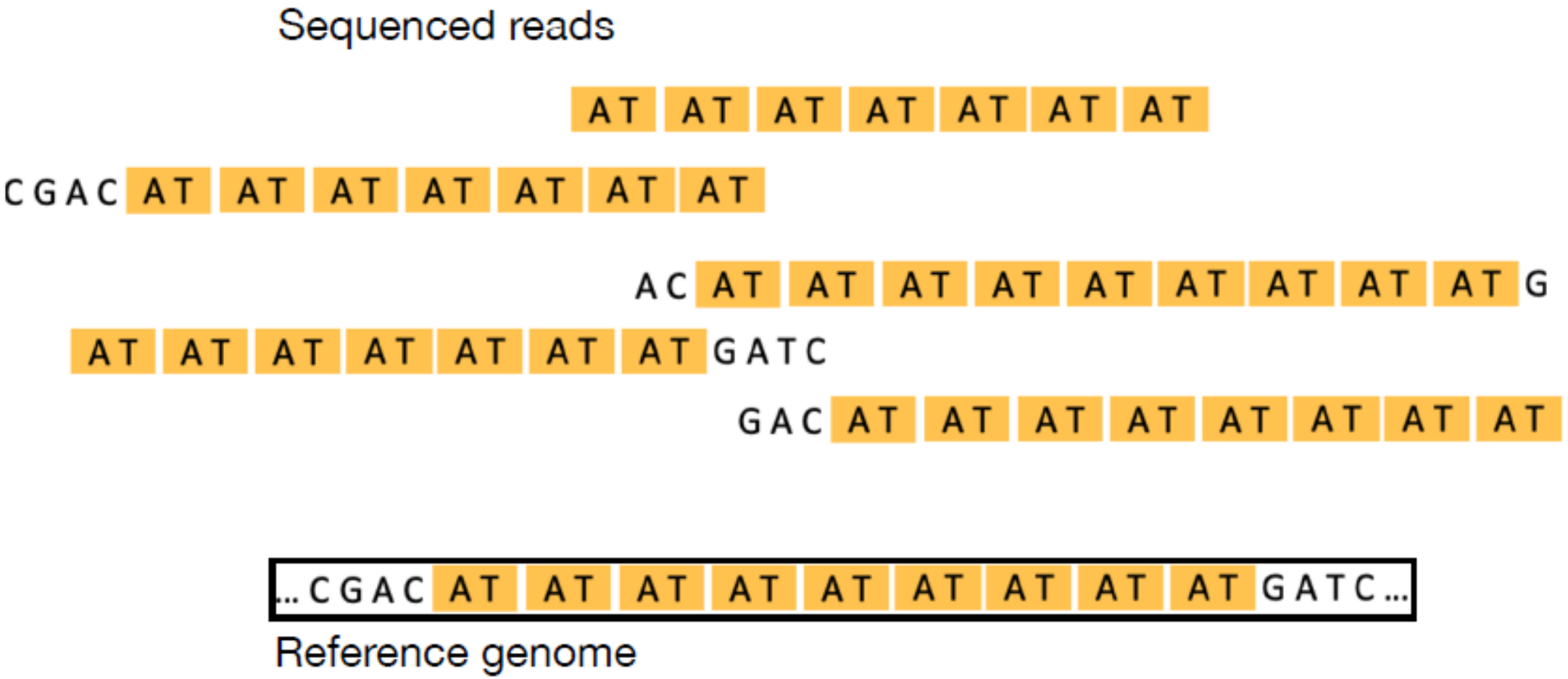
SNP:A->G

GTCTGGATGCT TCTATGCGGGCCCCT
GGTCTGGATGC TCTATGCGGGCCCC
CGGTCTGGATGC ATCTATGCGGGCC
GCGGTCTGGATG TATCTATGCGGGC
GCGGTCTGGAT TTATCTATGCGGG
GCGGTCTGGAT CTTATCTATGCGGG
GGCGGTCTGGAT CTTATCTATGCGG
GGCGGTCTGGA CTTATCTATGCGG

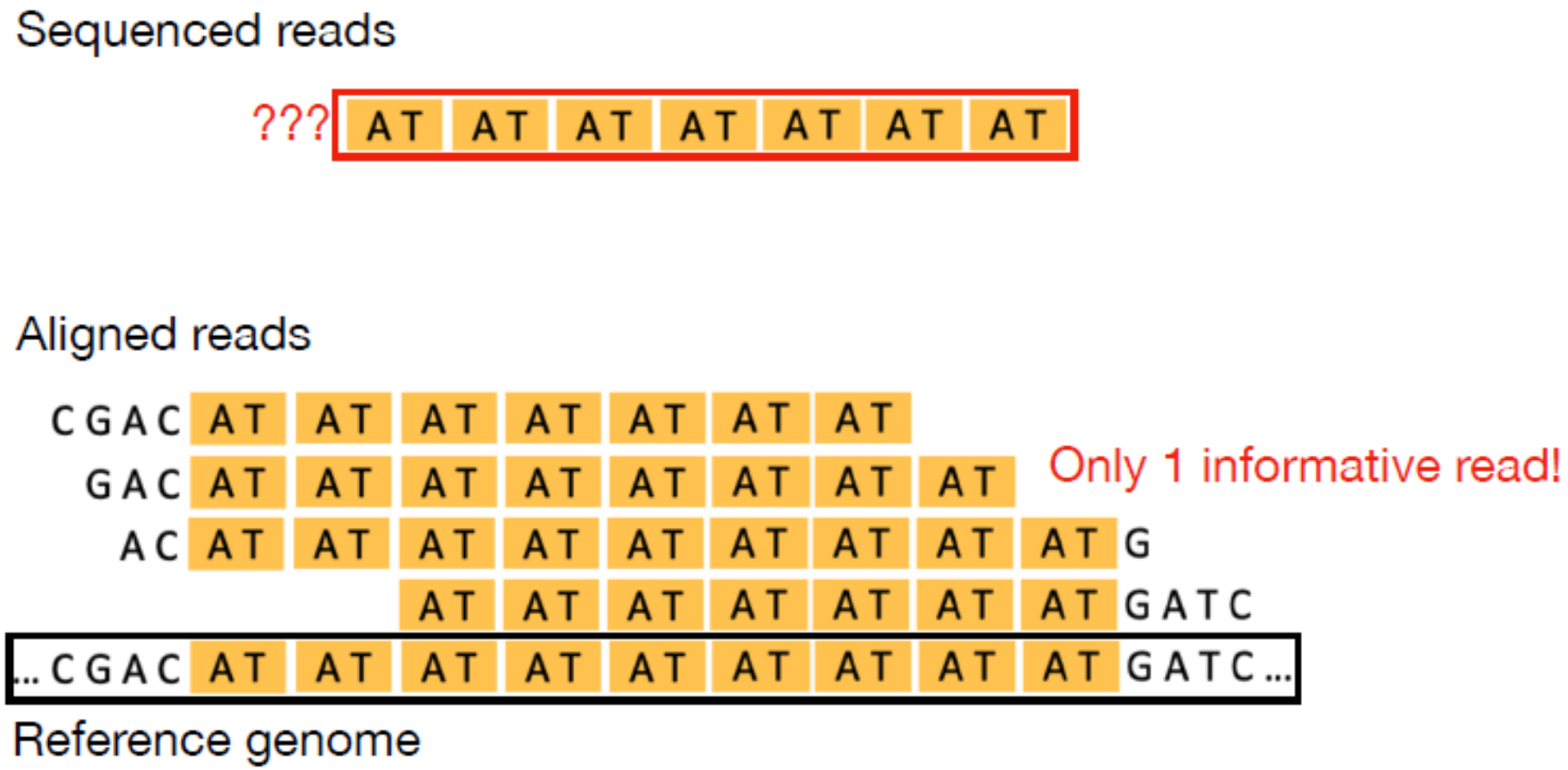
GGCGGTCTAGATGCTTATCTATGCGGGCCCCT

Reference genome sequence

Aligning short reads to STRs

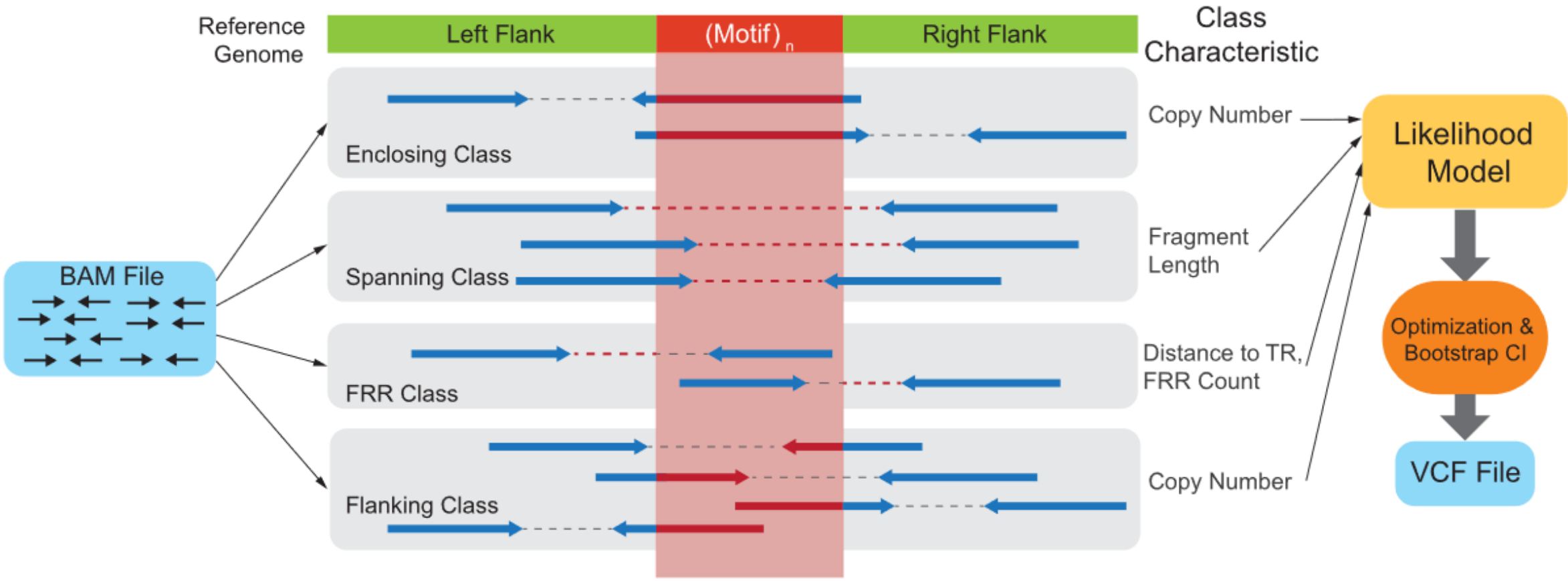


Aligning short reads to STRs



Specialised tools are needed

For example, GangSTR



References

[1] Fotsing, S.F., Margoliash, J., Wang, C., Saini, S., Yanicky, R., Shleizer-Burko, S., Goren, A. and Gymrek, M., 2019. The impact of short tandem repeat variation on gene expression. Nature genetics, 51(11), pp.1652-1659.

[2] Gymrek, M., 2017. A genomic view of short tandem repeats. Current opinion in genetics & development, 44, pp.9-16.

[3] Lundström, O.S., Verbiest, M.A., Xia, F., Jam, H.Z., Zlobec, I., Anisimova, M. and Gymrek, M., 2023. WebSTR: a population-wide database of short tandem repeat variation in humans. *Journal of molecular biology*, *435*(20), p.168260.

[4] Mousavi, N., Shleizer-Burko, S., Yanicky, R. and Gymrek, M., 2019. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic acids research, 47(15), pp.e90-e90.